ARTICLES

nature biotechnology

Genome sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*

Marco A van den Berg[1], Richard Albang[2], Kaj Albermann[2], Jonathan H Badger[3], Jean-Marc Daran[4,5], Arnold J M Driessen[4,6], Carlos Garcia-Estrada[7], Natalie D Fedorova[3], Diana M Harris[4,5], Wilbert H M Heijne[8], Vinita Joardar[3], Jan A K W Kiel[9], Andriy Kovalchuk[6], Juan F Martín[7,10], William C Nierman[3,11], Jeroen G Nijland[6], Jack T Pronk[4,5], Johannes A Roubos[8], Ida J van der Klei[4,9], Noël N M E van Peij[8], Marten Veenhuis[9], Hans von Döhren[12], Christian Wagner[2], Jennifer Wortman[3] & Roel A L Bovenberg[1]

Industrial penicillin production with the filamentous fungus *Penicillium chrysogenum* is based on an unprecedented effort in microbial strain improvement. To gain more insight into penicillin synthesis, we sequenced the 32.19 Mb genome of *P. chrysogenum* Wisconsin54-1255 and identified numerous genes responsible for key steps in penicillin production. DNA microarrays were used to compare the transcriptomes of the sequenced strain and a penicillinG high-producing strain, grown in the presence and absence of the side-chain precursor phenylacetic acid. Transcription of genes involved in biosynthesis of valine, cysteine and α-aminoadipic acid—precursors for penicillin biosynthesis—as well as of genes encoding microbody proteins, was increased in the high-producing strain. Some gene products were shown to be directly controlling β-lactam output. Many key cellular transport processes involving penicillins and intermediates remain to be characterized at the molecular level. Genes predicted to encode transporters were strongly overrepresented among the genes transcriptionally upregulated under conditions that stimulate penicillinG production, illustrating potential for future genomics-driven metabolic engineering.

Penicillins and derived β-lactam antibiotics have dramatically transformed health care and quality of life in the 80 years since Fleming's discovery of *Penicillium* that produces penicillins[1]. Large-scale production of β-lactam antibiotics is the result of sustained industrial strain improvement, representing numerous rounds of mutagenesis and selection. Although information on industrial processes is proprietary, product titers and productivities have increased by at least three orders of magnitude in the past 60 years[2], representing an unprecedented success in classical strain improvement.

Current industrial strains are derived from a single natural isolate of *P. chrysogenum*, NRRL1951, obtained during WWII from an infected cantaloupe[3]. Biochemical and genetic analysis of industrial strains led to the identification of several important mutations in high-producing strains, including amplification of penicillin biosynthesis genes[4]. However, much of the molecular basis for improved productivity remains to be elucidated. A detailed understanding of the molecular biology of *P. chrysogenum* is not only relevant for 'classical' penicillins. By applying genetic engineering approaches, it has become possible to extend the range of fermentation products to include β-lactam derivatives that could hitherto only be produced by chemical modification leading to great potential in terms of economy and sustainability. This is exemplified by the expression of the *Streptomyces clavuligerus cefE* gene, which encodes an expandase and has enabled high-yield production of cephalosporins with engineered *P. chrysogenum* strains[5].

Accessibility to the full range of genomics techniques will be invaluable for further innovation in antibiotics production. Here, we present the complete genome sequence of *P. chrysogenum* Wisconsin54-1255 (ref. 6). An *in silico* analysis of the genome sequence has focused on key processes in penicillin production. Moreover, DNA microarrays have been applied for transcriptome comparisons of the sequenced strain and a derived high-producing strain.

RESULTS

Genome sequence and analysis

The *P. chrysogenum* genome was sequenced by the whole-genome sequencing method. The nuclear genome of 32.19 Mb was covered by 49 supercontigs, including 21 supercontigs larger than 5 kb and

[1]DSM Anti Infectives, PO Box 425, 2600 AK Delft, The Netherlands. [2]Biomax Informatics AG, Lochhamer Str. 9, D- 82152 Martinsried, Germany. [3]The J. Craig Venter Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA. [4]Kluyver Centre for Genomics of Industrial Fermentation, Julianalaan 67, 2628 BC Delft, The Netherlands. [5]Department of Biotechnology, Delft University of Technology, Julianalaan 67, 2628 BC Delft, The Netherlands. [6]Molecular Microbiology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Kerklaan 30, 9751 NN Haren, The Netherlands. [7]INBIOTEC, Instituto de Biotecnología de León, Avda. Real no. 1, Parque Científico de León, 24006 León, Spain. [8]DSM Food Specialties, PO Box 1, 2600 MA Delft, The Netherlands. [9]Molecular Cell Biology, Groningen Biomolecular Sciences and Biotechnology Institute, Kerklaan 30, 9751 NN Haren, The Netherlands. [10]Area de Microbiologia, Departamento de Biologia Molecular, Facultad de CC. Biologicas y Ambientales, Universidad de Leon, Campus de Vegazana s/n 24071, Leon, Spain. [11]The George Washington University School of Medicine, Department of Biochemistry and Molecular Biology, 2300 Eye Street NW, Washington, DC 20037, USA. [12]Institut für Chemie, Universität Berlin, Sekretariat OE2, Franklinstrasse 29, 10623 Berlin, Germany. Correspondence should be addressed to M.A.v.d.B. (Marco.Berg-van-den@DSM.com).

Received 9 June; accepted 27 August; published online 28 September 2008; doi:10.1038/nbt.1498

Table 1 Genome statistics overview

Nuclear genome	
General information	
Size (Mb)	32.2
G+C content (mole%)	48.9
Coding (%)	56.6
Gene number (with similarities)	12,943 (11,472)
Putative pseudogenes	592
Truncated ORFs	116
Questionable ORFs	2
Genes (<100 aminoacids)	283
Mean gene length (bp)	1,515
Genes with intron (%)	10,812 (83.5)
Exons	
Mean number per gene	3
Mean length (bp)	434
GC content (mole%)	52.8
Introns	
Mean number per gene	2.2
Mean length (bp)	87.4
GC content (mole%)	45.3
Intergenic regions	
Mean length (bp)	842
GC content (mole%)	44.4
RNA	
tRNA number	145
5S rRNA number	28
Mitochondrial genome	
Size (bp)	31,790
GC content (mole%)	24.9
Gene number	17
Mean gene length	985.4
Coding (%)	52.7
Genes with intron (%)	0
tRNA number	26

14 supercontigs larger than 100 kb. Annotation based on a minimum open reading frame (ORF) size of 100 amino acids revealed 13,653 ORFs (**Table 1**), including 592 probable pseudogenes and 116 ORFs whose sequences were truncated because their coding regions spanned contig borders. Two ORFs were considered unlikely to encode proteins because of their small size, absence of detectable protein motifs and low codon adaptation index. The sequenced mitochondrial genome comprised 31,790 bp and 17 identified ORFs.

BLASTP matches were found for 11,472 ORFs ($P < 0.001$) to a nonredundant protein database, whereas the remaining 2,198 ORFs showed no significant similarities. Predicted protein-coding sequences account for 56.6% of the *P. chrysogenum* genome, with an average gene length of 1,515 bp. The GC content was 48.9% (52.8% for exons, 45.3% for introns and 44.4% for intergenic regions). On average, each gene contained 3.0 exons, with 83.5% of the genes containing introns. Using the FunCat classification system[7], 5,329 of the 12,943 predicted nuclear-encoded proteins could be assigned to the functional protein classes metabolism, energy, cellular transport and protein fate (**Fig. 1**).

Comparison with other fungal genomes

The sequenced *P. chrysogenum* genome is comparable in size to that of other filamentous fungi (**Supplementary Table 1** online). FunCat classification revealed a conserved orthologous core fungal proteome (**Supplementary Fig. 1** online) involved in energy production, protein fate and cell fate. Phylogenetic analysis based on the concatenated protein set (**Fig. 2a**) confirmed a close relationship to *Aspergillus* species. The tree topology indicated that *P. chrysogenum* is only distantly related to the other two sequenced *Penicillium* species, *Penicillium marneffei* and *Talaromyces stipitatus* (teleomorph of *Penicillium stipitatum*). This contradicts a previously published phylogeny[8] but is consistent with morphological observations[9].

The 14 largest supercontigs (between 167 kb and 6,387 kb), presumably correspond to chromosome arms or even entire chromosomes. Alignment against various Aspergilli chromosomes suggests extensive reshuffling has occurred after divergence of the *Aspergillus* and *Penicillium* lineages (**Fig. 2b**). Several supercontigs are bounded by areas with multiple synteny breaks, which may correspond to subtelomeric regions (**Fig. 2b**). Indications for subtelomeric instability have also been observed in Aspergilli[10] and *Magnaporthe oryzae*[11]. Five supercontigs (nos. 12, 16, 20, 21 and 22) contain gaps surrounded by larger syntenic blocks, which appear to be recombination cold spots. These gaps resemble putative centromeres in the eight *A. fumigatus* chromosomes[12]. Due to their high repeat content, centromeres as well as ribosomal DNA (rDNA) repeats regions typically do not get assembled into supercontigs in fungal genomes[12,13]. Coincidentally, the upper gap on supercontig no.16 is not surrounded by large syntenic blocks and is likely to contain the rDNA region (**Fig. 2b**).

Genome alignment revealed four supercontigs (nos. 17, 19, 23 and 24), representing 4% of the genome, that show little similarity to

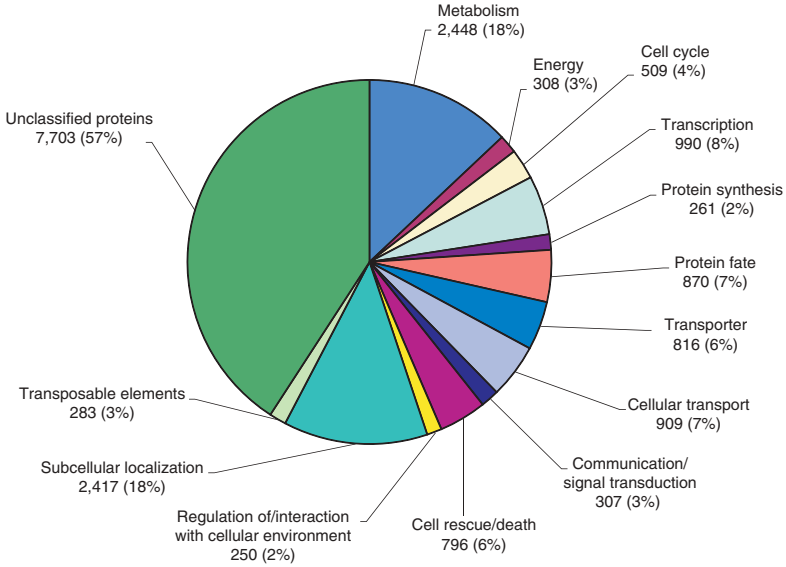


Figure 1 Genome characteristics of *P. chrysogenum*. Functional classification of *P. chrysogenum* ORFs. FunCat classes indicated are, 01 metabolism, 02 energy, 03 cell cycle and DNA processing, 04 transcription, 05 protein synthesis, 06 protein fate (folding, modification, destination), 67 transport facilitation, 08 cellular transport and transport mechanisms, 10 cellular communication/signal transduction mechanism, 11 cell rescue, defense and virulence, 13 regulation of/interaction with cellular environment, 40 subcellular localization, 29 transposable elements, 99 unclassified proteins.

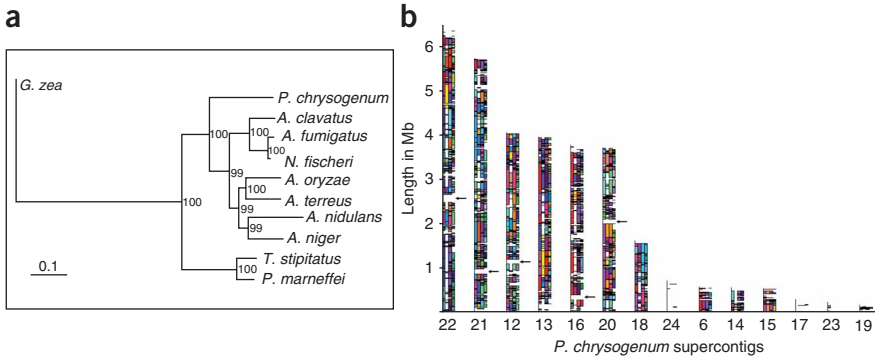


Figure 2 Comparison of the *P. chrysogenum* genome to other sequenced filamentous fungi. (**a**) Phylogenetic tree showing the relationships among sequenced *Aspergillus* and *Penicillium* species using *Gibberella zeae* as an outgroup. Branch lengths correspond to substitutions per site calculated using a maximum likelihood approach. Identical topologies were predicted using maximum parsimony and neighbor-joining methods. (**b**) Alignment of *A. nidulans, A. niger, A. fumigatus* and *A. oryzae* chromosomes against *P. chrysogenum* assemblies. Large assembled supercontigs (>100 kb) from four *Aspergillus* genomes were aligned to the 14 largest *P. chrysogenum* assemblies using MUMmer (http://mummer.sourceforge.net/). Regions with conserved gene order are represented by vertical columns of colored blocks. From left to right, *A. nidulans, A. niger, A. fumigatus* and *A. oryzae,* respectively. Each assembly (supercontig) from the target genomes is represented by a single color. Arrows indicate putative centromeres. Note: because the number of *A. nidulans* supercontigs (221) far exceeds the number of supercontigs available for the other Aspergilli, the length of aligned blocks between *A. nidulans* and *A. niger* may not represent the true extent of synteny between these two species.

Aspergilli. These regions contain *P. chrysogenum*-specific genes, which are typically smaller and contain fewer introns than other genes. Their biological roles are mostly unknown, although some seem to function in transport, metabolism or transcriptional regulation (**Supplementary Fig. 2c** online). These four nonsyntenic regions also contain numerous repeat elements and 23% of the genome's transposable elements (**Supplementary Table 2** online). Similar genomic islands have been found in other fungal genomes[11,13,14].

Almost 30% of the predicted *P. chrysogenum* proteins lack orthologs in other sequenced fungi. In the closely related genus *Aspergillus*, the origin of lineage-specific genes has been largely attributed to either gene acquisition through horizontal gene transfer[10,15] or to gene duplication followed by accelerated diversification and differential gene loss[13]. These genes tend to function in secondary metabolism and other accessory roles (**Supplementary Fig. 2a**) and may have a recent evolutionary origin. Phylogenetic analysis was applied to a subset of putative secondary metabolism genes. Thirty-three of such genes were identified using SMURF software (http://www.tigr.org/software/) encoding for: 20 polyketide synthases (PKS), 10 nonribosomal peptide synthetases (NRPS), 2 hybrid NRPS-PKS enzymes and 1 dimethylallyltryptophan synthase (**Supplementary Table 3** online). This is similar to the numbers found in Aspergilli[10,12,15,16]. The penicillin cluster is well known[17], and the siderophore synthetases for ferrichrome (Pc13g05250) and triacetylfusarinine (Pc16g03850, Pc22g20400) were readily assigned by homology (**Supplementary Table 4** online). None of the remaining six NRPS could be confidently identified. Pc21g15480 may encode roquefortine synthetase[18] and is clustered with tryptophan dimethylallyl transferase (Pc21g15430). The putative tetrapeptide synthetases Pc13g14330 and Pc16g04690 have similar architectures to those in Aspergilli and may form cyclopeptides with two adjacent D-amino acids presumably related to malformin[19]. Pc21g10790 may form a cyclohexapeptide containing a fatty-acid derived component and is orthologous to a similar NRPS found in *A. oryzae*.

Penicillin biosynthetic genes

Several prokaryotic features of two penicillin biosynthetic genes, *pcbAB* and *pcbC*, encoding α-aminoadipoyl-L-cysteinyl-D-valine synthetase and isopenicillinN (IPN) synthase, suggested that the penicillin gene cluster emerged through horizontal gene transfer from bacteria to fungi[20]. Both genes lack introns (which is unique for large NRPS genes like *pcbAB*), are highly homologous to their bacterial counterparts and are physically linked. Other features to consider are GC content (which is above 60% in prokaryotic penicillin producers) and specific codon usage. In *P. chrysogenum*, the GC content of the penicillin biosynthetic genes is only slightly higher than the overall genome average (**Supplementary Data** online). In the clavulanic acid producer *S. clavuligerus* the phenylalanine-codon UUU is extremely rare compared to UUG; UUU comprises only ~2.3% of total phenylalanine-codons. Whereas, in *P. chrysogenum*, the UUU codon overall is used in one-third of the cases, it is used for 26.2% and 17.6% of the phenylalanine-codons in *pcbAB* and *pcbC*, respectively. This can be interpreted as near complete codon adaptation because of the hypothesized transfer acquisition event.

Three other examples of possible horizontal gene transfer were identified in the *P. chrysogenum* genome: the arsenate-resistance cluster and two 6-methylsalicylic acid clusters (**Supplementary Data** and **Supplementary Table 4**). These gene clusters contain highly conserved bacterial-like genes with GC content well above the surrounding genes (exons with 55–58% GC).

The penicillin biosynthetic genes are clustered on supercontig 21 in the middle of a 120-kb region that is amplified in industrial *P. chrysogenum* strains[4]. Thirty-nine additional ORFs were identified in this region (**Supplementary Table 5** online), including genes encoding transporters and transcriptional regulators. However, the predicted annotation of these ORFs does not suggest clear functions in penicillin biosynthesis as reported recently[21,22].

The third penicillin biosynthetic gene, *penDE,* encoding acyl-CoA: isopenicillinN acyltransferase, has a paralog, Pc13g09140. This gene was not transcribed under the conditions studied (**Supplementary**

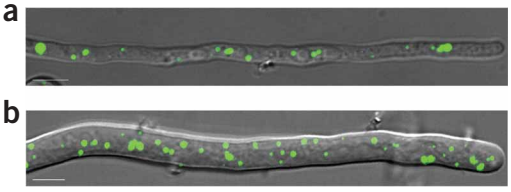


Figure 3 Microbodies in *P. chrysogenum* strains. (**a**,**b**) Cells of the *P. chrysogenum* type strain NRRL1951 (**a**) and the high penicillin producing strain DS17690 (**b**) producing the microbody-targeted protein GFP. SKL (**Supplementary Data**) were cultivated in batch cultures on penicillin-production media. Hyphae of strain DS17690 show enhanced microbodies numbers relative to the type strain NRRL1951. Scale bar, 5 μm.

Table 6 online). Detailed analyses are needed to reveal its actual function. Several orthologs of β-lactam biosynthesis genes were identified throughout the genome (**Supplementary Table 6**). As deletion of *phl*, which encodes phenylacetyl-CoA ligase, resulted only in a partial loss of penicillinG production[23], other phenylacetyl-CoA ligases must be present[24]. These may include identified orthologs of 4-coumarate-CoA ligase. Surprisingly, several orthologs of bacterial[25] and fungal[26] isopenicillinN epimerase were identified. The predicted protein sequence of Pc12g11540 shares 40% homology with *S. clavuligerus* isopenicillinN epimerase, although it probably functions as an aminotransferase. Also, orthologs of *Acremonium chrysogenum cefD1* and *cefD2* were identified. The presence of these ORFs is remarkable, as *P. chrysogenum* can only produce penicillinN after introduction of both *A. chrysogenum* genes[27]. The *P. chrysogenum* ORFs may be remnants of an ancestral cephalosporin pathway.

Microbodies

In *P. chrysogenum,* microbodies (peroxisomes) are essential for penicillin biosynthesis because the two final enzymatic steps catalyzed by acyl-CoA:isopenicillinN acyltransferase[28] and phenylacetyl-CoA ligase[29] are located in these organelles. Moreover, high-producing strains have enhanced microbody volume fractions (**Fig. 3**). Also, a further increase in microbody abundance by overexpression of the proliferation gene *pex11* leads to a significant increase in penicillin production[30]. Genome 2D-searches[31] with known consensus sequences for microbody targeting signals (PTS)[32] identified 214 putative matrix proteins (196 and 17 with putative PTS1 and PTS2 respectively; 1 with both signals) (**Supplementary Table 7** online). Remarkably, the putative isopenicillinN-CoA epimerase (Pc22g13680) has a predicted PTS1. Many of the proteins are β-oxidation homologs, including multiple acyl-CoA synthetases and putative 3-ketoacyl-CoA

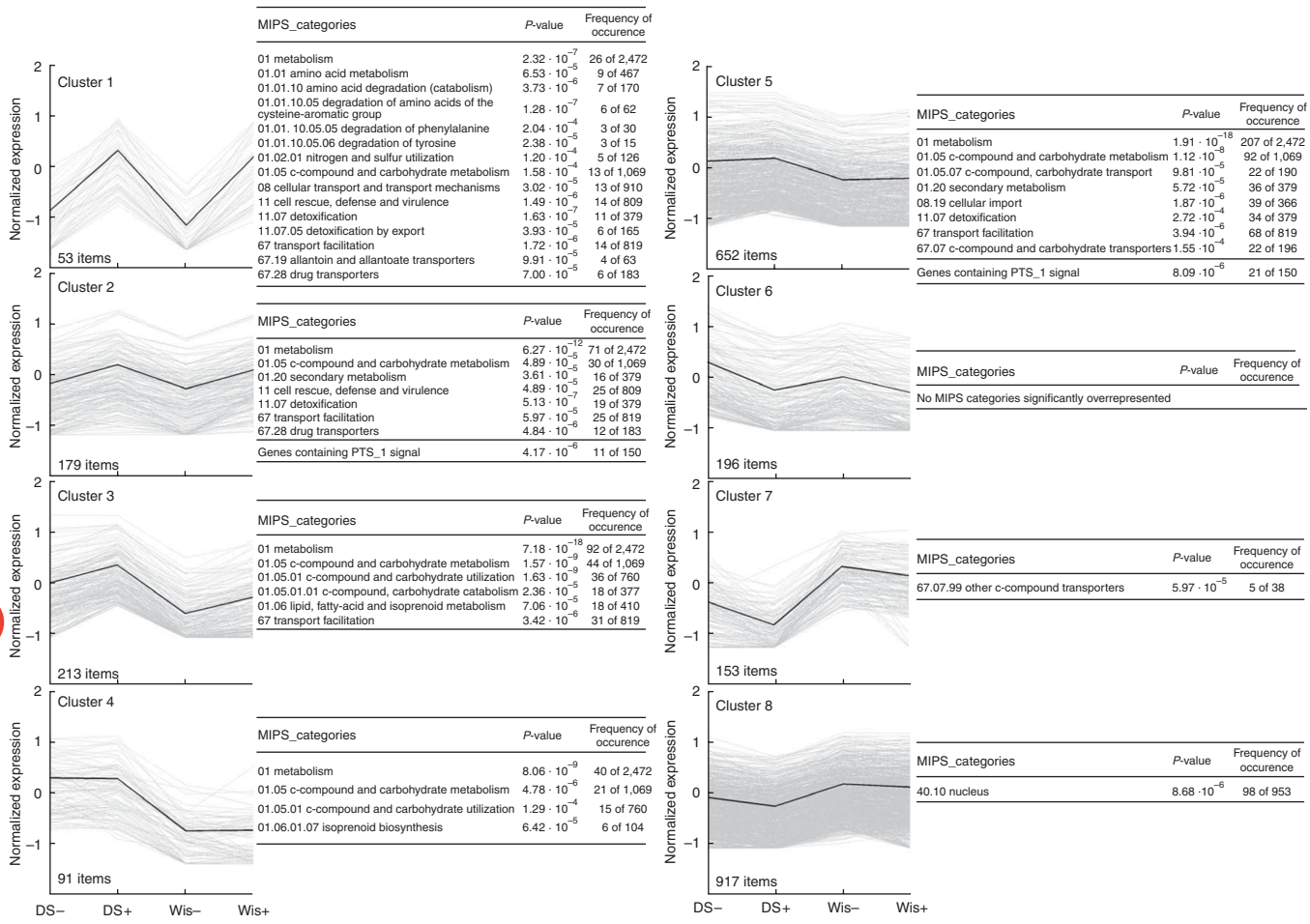


Figure 4 Transcriptional response of significantly changed genes, *K*-means clustered and the overrepresented functional categories in each cluster. Genes that showed a significantly different transcript level in at least one comparison (DS17690 + (DS+) versus – PAA (DS–); Wisconsin54-1255 + (Wis+) versus – PAA (Wis–); DS17690 + PAA versus Wisconsin54-1255 + PAA and DS17690 – PAA versus Wisconsin541255 – PAA) were grouped in eight clusters by *K*-means clustering. The thick lines represent the average of the mean normalized transcript levels of the genes in each cluster. The *y* axis represents ^{10}log transcript levels. Three main categories were observed; (i) genes whose transcript levels were not influenced by the strain improvement program but with a higher transcript level in the presence of PAA (clusters 1 and 2); (ii) genes with a constitutively higher transcript level in DS17690 than in Wisconsin54-1255, irrespective of the presence of PAA (clusters 4 and 5); (iii) genes with a lower transcript level in DS17690, irrespective of the presence of PAA (clusters 7 and 8). Cluster 3 contains genes that showed a higher transcript level in DS17690 than in Wisconsin54-1255 and only responded to PAA in DS17690. Cluster 6 contains genes that only showed a lower transcript level in the presence of PAA in the DS17690 strain. Functional categories are mentioned together with a *P*-value indicating their overrepresentation in each cluster and the number of genes belonging to functional category in the cluster relative to the total number of these genes in the genome. Due to redundancy in the functional categories, statistically significant overrepresentation may not always reflect biological significance. Genes encoding enzymes containing the peroxisomal targeting signal PTS1 occurring in each cluster are indicated in the same manner.

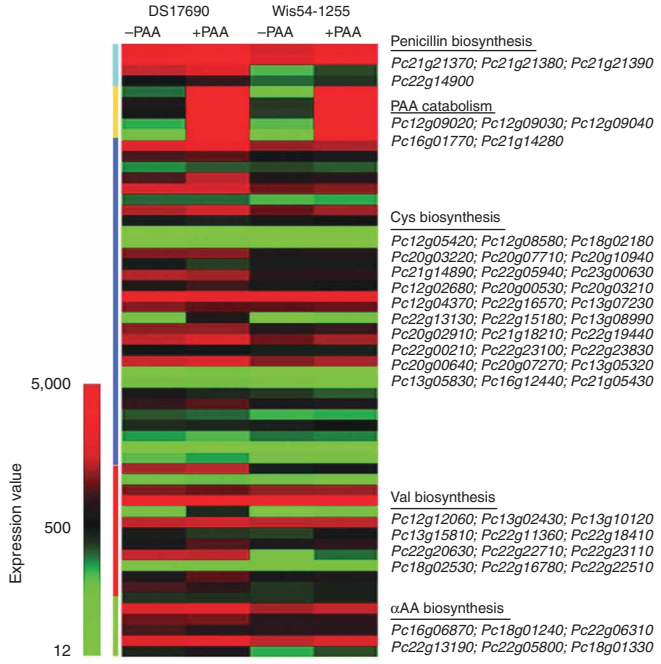


Figure 5 Eisen diagram of transcript levels of key genes in penicillin biosynthesis, amino-acid biosynthesis and phenylacetate catabolism. *P. chrysogenum* Wisconsin54-1255 and DS17690 strains were grown in glucose-limited chemostat cultures in the absence and presence of phenylacetic acid (PAA). The color bar indicates the range of the transcript levels for each gene, normalized to the average transcript level of the gene. Genes (putatively) related to penicillin biosynthesis, synthesis of the penicillin precursors cysteine, valine and α-aminoadipic acid and to phenylacetate catabolism are depicted.

thiolases (Pc13g12930, Pc15g00410 and Pc22g06820). Indeed, *P. chrysogenum* readily consumes oleate as the sole source of carbon and energy. Other PTS-containing proteins, such as D-amino acid oxidases, may play a role in the metabolism of various carbon or nitrogen sources.

Transcriptome analysis

When genomic DNA of *P. chrysogenum* was hybridized to microarrays, 99.4% of the probe sets hybridized (**Supplementary Data**). Once validated (**Supplementary Data**), these microarrays were used to investigate, at the transcriptome level, the molecular basis of the improved penicillin productivity achieved via classical strain improvement. Transcriptome analysis was performed on aerobic, glucose-limited chemostat cultures of Wisconsin54-1255 and the derived industrial, high-producing strain DS17690 (ref. 33). Some intermediates of β-lactam biosynthesis are produced in the absence of the side-chain phenylacetic acid (PAA)[33], but penicillinG biosynthesis is strictly dependent on PAA[34] (**Supplementary Data**). In DS17690 grown in the absence of PAA, 67% of the genome (~9,200 genes) yielded a detectable transcript (**Supplementary Table 8** online). Under comparable conditions, 86% (~5500 genes) of the smaller *S. cerevisiae* genome was transcribed[35]. To discriminate between PAA-responsive transcripts and transcripts potentially related to improved penicillinG productivity, the two strains were grown under penicillinG-producing and nonproducing conditions. In at least one of the four comparisons, 2,470 genes were differentially transcribed (**Supplementary Table 8**). By *K*-means clustering[36], these genes were assigned to eight clusters (**Fig. 4** and **Supplementary Tables 9–16** online).

Transcription of the penicillinG biosynthesis genes *pcbAB*, *pcbC* and *phl* was independent of PAA, but two- to fourfold higher in the high-producing strain (**Fig. 4**, cluster 5). *penDE* showed a similar trend (1.9- and 1.5-fold difference in the presence and absence of PAA, respectively; $P < 0.05$ in a *t*-test). Genes encoding enzymes involved in the biosynthesis of the amino-acid precursors of penicillin (cysteine, valine and α-aminoadipic acid) were also transcribed at higher levels in the high-producing strain independent of the presence of PAA (**Fig. 4**, cluster 5 and **Fig. 5**). This included sulfur reduction and early stages of serine (and cysteine) biosynthesis (Pc20g03220; Pc12g02680 and Pc12g04370), as well as a homolog of *O*-acetyl-homoserine (thiol)-lyase (Pc12g05420), a key enzyme in the trans-sulfuration pathway toward cysteine. Several genes encoding enzymes related to α-aminoadipic acid (lysine; Pc18g01330, Pc14g00150) and valine (Pc22g22510, Pc22g23110) metabolism showed a similar trend.

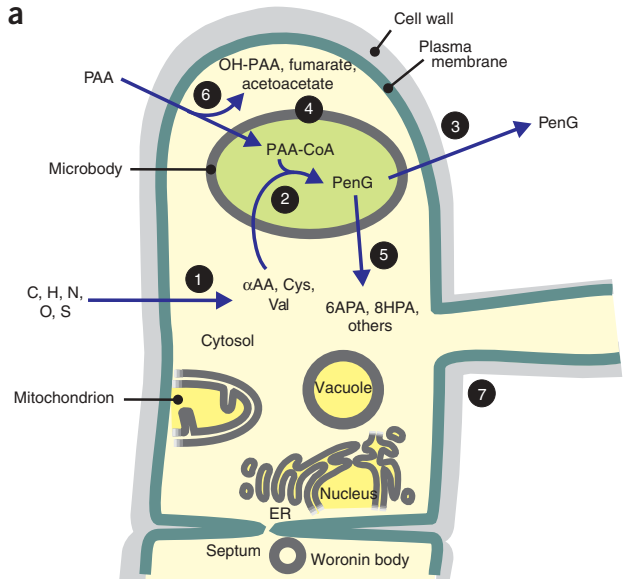
Of the genes predicted to encode microbody proteins, 27 showed higher transcript levels in DS17690, irrespective of PAA addition (**Fig. 4**, cluster 5). This class was also overrepresented among genes that were upregulated by addition of PAA (**Fig. 4**, cluster 2).

The homogentisate pathway for PAA degradation has been reported to be largely inactivated in Wisconsin54-1255 and, presumably, also in derived strains, owing to point mutations in the *pahA* gene encoding phenylacetate hydroxylase[37]. Nevertheless, both strains showed very low, but significant rates of PAA consumption that could not be attributed to penicillinG production (**Supplementary Data**). Despite the low *in vivo* activity of this homogentisate pathway, its transcriptional regulation has been retained throughout the strain improvement program, as its structural genes showed increased transcript levels in the presence of PAA in both strains (**Fig. 4**, clusters 1 and 2).

Several transcriptional regulators have been implicated in the transcription of *pcbAB*, *pcbC* and *penDE* in *P. chrysogenum*[38]. No penicillin-specific transcriptional regulator has been identified, although strong effects were reported from an enhancer sequence in the upstream region of *pcbAB*[39] and chromatin modulation[40]. The *laeA* gene responsible for the latter effect strongly affects secondary metabolism in *Aspergillus fumigatus*[40]. Although the *P. chrysogenum* ortholog was transcribed, its transcript levels were not substantially influenced by strain or cultivation conditions (Pc16g14010, **Supplementary Table 17** online). Several other putative transcription factors, whose functions remain to be elucidated, were found to be associated with secondary metabolite clusters (**Supplementary Data**).

Transport

Transport mechanisms for β-lactam antibiotics and intermediates across the fungal plasma membrane and between intracellular compartments are poorly understood. Industrial fermentations yield high amounts of penicillinG in the external broth, whereas intracellular concentrations are typically tenfold lower. Moreover, penicillinG secretion is sensitive to verapamil[41], an antagonist of multidrug transporters. This implies that secretion is an active process possibly mediated by (an) ABC transporter(s), whose identity has remained elusive. *P. chrysogenum* contains 830 genes that specify transporter proteins. Secondary transporters (688) are numerous with the majority belonging to the major facilitator superfamily (416), whereas 51 ABC transporters were identified. The functional categories metabolism, transport and detoxification were among the most strongly overrepresented in the gene clusters that were transcriptionally upregulated in the presence of PAA in both strains (**Fig. 4**, clusters 1 and 2 and **Supplementary Table 18** online). Several of these showed



b

	Description	No. of genes higher expressed	No. of genes lower expressed
1	Growth and metabolism	106	75
2	From precursors to penicllinG	6	1
3	PenicillinG secretion	13	4
4	Peroxisome related	2	4
5	β-lactam metabolism	4	2
6	PAA metabolism	6	0
7	Cell differentiation and morphology	14	11
8	Other	156	174
	Subtotal	307	271

Figure 6 Categorization of differentially transcribed genes in DS17690 and Wisconsin54-1255. (**a**) Schematic representation of a *P. chrysogenum* cell and eight functional categories of cell metabolism directly influencing penicillin productivity: 1, growth and metabolism (from nutrients to precursors); 2, from precursors to penicillinG; 3, penicillinG secretion; 4, peroxisome related; 5, β-lactam metabolism; 6, PAA metabolism; 7, cell differentiation and morphology; 8, other (including unknowns, cell signaling and communication, transcription factors, stress response). (**b**) Transcript levels of all genes in DS17690 and Wisconsin54-1255 were compared under producing conditions (+PAA). Significant changes were selected by statistical analysis of microarrays (SAM) analysis with a threshold fold-change of 2 and a false-discovery rate of $<1\%$, thus identifying 1,605 genes. An additional, more stringent criteria were imposed to select only genes whose highest average transcript level signal (in either strain) was larger than 200. The 578 remaining genes were functionally categorized according to the eight categories.

sequence similarity to known multidrug transporter genes. Transporters were significantly overrepresented within the class of genes expressed more highly in DS17690 than in Wisconsin54-1255, irrespective of PAA ($P = 3.94 \times 10^{-6}$, see **Fig. 4**, cluster 5), identifying sixty-eight potential active transporters (**Supplementary Table 19** online). Interestingly, none of the previously suggested penicillin transporters[41,42] are among this group. Although some transporter genes may be involved in transport of PAA rather than of β-lactams or intermediates, the strong enrichment of transporter genes suggest that penicillin secretion might result from the simultaneous activity of multiple transporters.

DISCUSSION

Although observations on amplification of penicillin biosynthetic genes[4] and increased abundance of microbodies[30] have provided some insights, the question of how strain improvement transformed *P. chrysogenum* into an efficient penicillin producer has remained a formidable challenge. Availability of a complete genome sequence enables an integral approach. The potential is illustrated by several results, including increased transcript levels in a high-producing strain of genes involved in side-chain activation, of α-aminoadipate, valine and cysteine biosynthesis genes, and of genes encoding multidrug transporters. Three hundred and seven of these upregulated genes were assigned to 8 categories directly influencing penicillin productivity (**Fig. 6**). A similar number of genes (271) showed reduced transcript levels in the high-producing strain (**Fig. 6**). Gene-by-gene analyses of this latter set suggests an overrepresentation of genes encoding morphogenesis and developmental factors. This is in line with the morphological differences between the two strains. These results indicate that metabolic control of penicillin production in high-producing strains is likely to be distributed over many cellular processes that have been iteratively optimized in the long strain improvement history. Scattered gene duplications and deletions throughout the genome of different strains (**Supplementary Data**) may contribute to this combined up- and downregulation of different pathways and processes. However, most of these genes encode hypothetical proteins, except for Pc12g16490 (**Supplementary Table 20** online, cluster no. 4), involved in cell development.

Despite the massive improvements already achieved in classical strain improvement, our results indicate that further improvement of penicillin production remains a possibility. For instance, the high-level expression of the PAA degradation pathway may impose an energetic burden and thus affect productivity. Moreover, the genome sequence shows evidence for other, potentially competing, β-lactam pathways that may contribute to by-product formation. Some of the genes, which strongly resemble genes encoding isopenicillinN epimerase and acetyltransferase, are highly transcribed and show increased levels in the high-producing strain DS17690. Further improvements can be anticipated from a detailed analysis and subsequent targeted modification of genes involved in microbody biogenesis and function, and of membrane transporters involved in the (intra)cellular transport of products and intermediates.

When a subset of the genes that showed altered transcript levels in DS17690 relative to Wisconsin54-1255 were knocked out, several amino-acid biosynthesis genes that were transcriptionally upregulated were confirmed to be involved in the increased β-lactam productivity (**Supplementary Table 21** online). This is in contrast to the gene encoding Pc20g0240, a putative transporter with high mRNA levels in DS17690, which, when silenced, has little effect on β-lactam biosynthesis. Surprisingly, the knockout of a putative isopenicillinN-CoA epimerase seems to lead to a small increase in β-lactam productivity. In *Cephalosporium acremonium* this enzyme is involved in IPN to penicillinN conversion[27]. It is tempting to speculate that this enzyme might be involved in a so far unidentified side reaction in *P. chrysogenum* reducing the overall yield of secreted penicillinG. These preliminary and diverse results demonstrate the need for a thorough follow-up of the transcriptome-based identification of targets for metabolic engineering.

The availability of a highly reproducible platform for transcriptome analysis will facilitate the systematic analysis of *P. chrysogenum* strain lineages, with the aim to further elucidate the molecular basis for the substantial increase in penicillinG productivity that has occurred during six decades of classical strain improvement. In addition, a full exploitation of the *P. chrysogenum* genome sequence will require the integration of additional levels of cellular information (for example, metabolome and proteome), as well as the construction of

genome-scale metabolic models[43]. Such 'systems' approaches will ultimately contribute to further improvement of this important cell factory via inverse metabolic engineering[44].

METHODS

Genome sequencing and assembly. The genome of *P. chrysogenum* Wisconsin54-1255 (ATCC28089) was sequenced by the whole-genome random sequencing method[12] by obtaining paired-end reads from five libraries. Four plasmid libraries with inserts prepared by random shearing were constructed with insert sizes of 3–4 kb, 4–6 kb, 6–12 kb and 25–50 kb. Approximately 2.3× genome coverage was obtained from each of these libraries. In addition, 16,000 sequences were obtained from the ends of BAC clones. The BAC libraries were prepared by partial digestion of the *P. chrysogenum* genomic DNA using either *Bam*HI or *Hin*dIII[45]. The mean insert size of these two libraries was 130 kb. Assembly of these sequences generated 49 supercontigs, with 14 of these larger than 100 kb. The sizes of the largest five supercontigs are 6.2 Mb, 5.6 Mb, 4.0 Mb, 3.9 Mb and 3.6 Mb. The final genome coverage in contigs was 9.8×. Genome assembly was accomplished using Celera Assembler[46]. The mitochondrial sequence was obtained by manual review of the assembled contigs by sequence similarity to the mitochondrial sequence of *A. fumigatus* Af293. The genome sequence was deposited at EMBL under the accession numbers AM920416-a.m.920464.

Genome annotation and analysis. Analysis and annotation of the genomic sequences of *P. chrysogenum* was performed with a combined automatic and manual approach. For all supercontigs larger than 5 kb ORFs were predicted by a version of FGENESH[47] trained on sequences of Ascomycetes, as well as other algorithms (**Supplementary Methods** online). ORFs were named after the organism (Pc), supercontig number (two digits) followed by g (gene) and a five-digit number matching the order of the ORFs on the contig. For all ORFs identified from the above described approach an exhaustive automatic bioinformatic analysis in respect to function and structure of the respective protein was performed using the PEDANT-Pro^{T} software[48]. Annotation of description, functional categories according to the Functional Catalog (FunCat) classification system[7], and Enzyme Commission (EC) numbers have been performed for each *P. chrysogenum* ORF with a multi-step semiautomatic approach (**Supplementary Methods**). tRNA genes were identified as described in **Supplementary Methods**.

Repetitive elements. Identification of repeat elements was performed using RepeatMasker (http://www.repeatmasker.org/), RepeatScout (http://repeatscout.bioprojects.org/), Tandem Repeats Finder (http://tandem.bu.edu/trf/trf.html) and Transposon-PSI (Brian Haas, http://transposonpsi.sourceforge.net). The result of the latter was used to calculate the densities as the percentage of nucleotide bases in the regions of interest that overlap with repeat or transposable elements (**Supplementary Table 2**).

Genome-genome alignment. Pair-wise alignments between *P. chrysogenum* and four target genomes (*A. fumigatus*, *A. oryzae*, *A. niger* and *A. nidulans*) were performed using the Promer program of the MUMmer package (http://mummer.sourceforge.net/). Supercontigs larger than 100 kb of the four *Aspergillus* species were aligned against the 14 largest *P. chrysogenum* supercontigs.

Phylogenetic analysis and species tree. To generate the species tree, a total of 90 orthologous genes from *P. chrysogenum*, *P. marneffei* (GenBank ABAR00000000), *T. stipitatus* (GenBank ABAS00000000), *A. niger*[16], *A. nidulans*[10], *A. oryzae*[15], *A. fumigatus*[12], *A. clavatus* (GenBank AAKD00000000), *A. terreus* (Refseq NT_165925-NT_165951), *Neosartorya fischeri* (GenBank AAKE00000000) and *Gibberella zeae* (anamorph *Fusarium graminearum*)[49] were aligned at the amino acid level (**Supplementary Table 22** online). To minimize the effect of incorrect or incongruent gene models, these proteins were chosen on the basis of having identical numbers of introns in each species and similar lengths (95% overlap). Sequences were aligned using Muscle[50]. DNA alignments were then concatenated and passed to the Phylip package. Maximum likelihood trees were calculated on each replicate, applying the JTT substitution model with a gamma distribution (alpha = 0.5) of rates over four categories of variable sites, and a consensus tree was produced.

Chemostat cultivations. Independent triplicate chemostat cultures of *P. chrysogenum* Wisconsin54-1255 and the high-producing penicillinG strain DS17690 were run in the presence and absence of phenylacetic acid exactly as described before[33].

Microarray methods (detailed procedures are available in **Supplementary Methods**). The *P. chrysogenum* genome sequence information was used to prepare a proprietary DNA microarray, using the Affymetrix Custom GeneChip program (Affymetrix): GeneChip, DSM_PENa520255F. Samples from chemostat cultures were filtered within seconds and quenched in liquid nitrogen. Total RNA isolation with Trizol reagent and acid phenol-chloroform for extraction was followed by cDNA synthesis and cRNA synthesis. Hybridized arrays were scanned and analyzed using the Affymetrix GeneChip Operating Software (GCOS, Affymetrix). All arrays were globally scaled to a target value of 100 using the average signal from all gene features using GCOS. Differential expression was assessed using the Significance Analysis of Micro arrays (SAM version 1.21) add-in to Microsoft Excel[34]. The fold-change threshold and the false discovery rate values were set at 2% and 1%, respectively. Enrichment of Munich Information Center for Protein Sequences (MIPS) categories was assessed by Fisher's Exact test employing hypergeometric distribution with a *P*-value cut-off of $3{\cdot}10^{-4}$ (with a Bonferroni correction). The genes with significantly changed expression in one of the comparisons were arranged in clusters by the *K*-means clustering tool of Genedata Expressionist (Genedata). The coefficient of variation (CV) of transcript levels of independent, triplicate chemostat cultures did not exceed 21% (**Supplementary Data**) and was similar to that of transcriptome analysis on chemostat cultures of the nonfilamentous fungus *S. cerevisiae*[34].

Accession number. The array data were deposited at Genome Expression omnibus under the serial number GSE9825. The genome sequence was deposited at EMBL under the accession numbers AM920416-AM920464.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

J.-M.D. and J.T.P. wish to acknowledge the financial support from The Netherlands Genomics Initiative. C.G.-E. is supported by the Torres Quevedo Program (PTQ04-3-0411). D.M.H. and J.G.N. are supported by The Netherlands Organisation for Scientific Research (NWO) via the IBOS Programme (Integration of Biosynthesis and Organic Synthesis) of Advanced Chemical Technologies for Sustainability (ACTS). J.A.K.W.K. and A.K. are financially supported by The Netherlands Ministry of Economic Affairs and the B-Basic partner organizations (www.b-basic.nl) through B-Basic, a public-private NWO-ACTS program. J.F.M. is supported by a grant of the European Union (EUROFUNGBASE – FP6-018964). Part of this work was supported by SENTER (IS project ISO43055). Jiaqi Hunag, Resham Kulkarni and Charles Lu worked on the initial auto-annotation. Dave Gaiser assisted with the phylogenetic analysis. Zita van der Krogt is acknowledged for performing the chemostat cultivations and Mareike Viebahn for her contribution to the transcriptome analyses. We thank Wieb H. Meijer for fluorescence microscopy. Theo Knijnenburg is thanked for performing the hypergeometric distribution analyses. Hilde Huininga and Jasper Deuring are acknowledged for isolating DNA. Hilly Menke is acknowledged for facilitating the gDNA comparisons. Hesselien Touw, Linda van den Hoogen, Hilde Huininga and Laurens Ekkelkamp made the knockout mutants. Dick Schipper was responsible for β-lactam analyses.

AUTHOR CONTRIBUTIONS

M.A.v.d.B. initiated and coordinated the *P. chrysogenum* genome project. W.C.N. coordinated the genome sequencing. N.D.F., J.H.B., V.J. and J.W. were responsible for genome assembly and genome-genome comparisons. K.A., R.A and C.W. were responsible for genome annotation and analysis. C.G.-E. and J.F.M. performed the analysis of transcription factors. H.v.D. and N.D.F. analysed the secondary metabolites and gene clusters. J.A.K.W.K., I.J.v.d.K. and M.V. performed the analysis of the microbody proteins and induction. A.K., J.G.N. and A.J.M.D. carried out the transporter analysis. N.N.M.E.v.P. was responsible for the microarray design in collaboration with Affymetrix. W.H.M.H. and J.A.R. performed MicroArray data analyses. D.M.H., J.T.P and

J.-M.D. were responsible for chemostat cultivations and subsequent MicroArray data analyses. J.G.N. performed the analysis of MFS and A.K. did the analysis of ABC transporter clusters. J.T.P., A.J.M.D., W.C.N., R.A.L.B. and M.A.v.d.B. wrote the final text of the manuscript.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Fleming, A. The antibacterial action of a *Penicillium*, with special reference to their use in the isolation of *B. influenzae*. *Br. J. Exp. Pathol.* **10**, 226–236 (1929).
2. Hersbach, G.J.M., Van der Beek, C.P. & Van Dijck, P.W.M. The penicillins: properties, biosynthesis, and fermentation. in *Biotechnology of Industrial Antibiotics. Drugs and the Pharmaceutical Sciences* vol. 22 (ed. Vandamme, E.J.) 45–140, (Marcel Dekker, New York, 1984).
3. Raper, K.B., Alexander, D.R. & Coghill, R.D. Penicillin. II. Natural variation and penicillin production in *Penicillium notatum* and allied species. *J. Bacteriol.* **48**, 639–659 (1944).
4. Fierro, F. *et al.* The penicillin biosynthetic gene cluster is amplified in tandem repeats linked by conserved hexanucleotide sequences. *Proc. Natl. Acad. Sci. USA* **92**, 6200–6204 (1995).
5. Cantwell, C.A. *et al.* Cloning and expression of a hybrid *Streptomyces clavuligerus cefE* gene in *Penicillium chrysogenum*. *Curr. Genet.* **17**, 213–221 (1990).
6. Elander, R. Strain improvement and preservation of beta-lactam producing microorganisms. in *Antibiotics Containing the Beta-Lactam Structure I*. (eds. Demain, A.L. & Solomon, N.) 97–146, (Springer-Verlag, New York, 1983).
7. Ruepp, A. *et al.* The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.* **14**, 5539–5545 (2004).
8. Verweij, P.E. *et al.* Phylogenetic relationships of five species of *Aspergillus* and related taxa as deduced by comparison of sequences of small subunit ribosomal RNA. *J. Med. Vet. Mycol.* **33**, 185–190 (1995).
9. Malloch, D. The Trichomaceae: relationships with other Ascomycetes. in *Advances in Penicillium and Aspergillus Systematics* (eds. Samson, R.A. & Pitt, J.I.) 365–382, (Plenum Press, New York, 1985).
10. Galagan, J.E. *et al.* Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* **438**, 1105–1115 (2005).
11. Rehmeyer, C. *et al.* Organization of chromosome ends in the rice blast fungus, *Magnaporthe oryzae*. *Nucleic Acids Res.* **34**, 4685–4701 (2006).
12. Nierman, W.C. *et al.* Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature* **438**, 1151–1156 (2005).
13. Thon, M.R. *et al.* The role of transposable element clusters in genome evolution and loss of synteny in the rice blast fungus *Magnaporthe oryzae*. *Genome Biol.* **7**, R16 (2006).
14. Fedorova, N.D. *et al.* Genomic Islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. *PLoS Genet.* **4**, 1000046 (2008).
15. Machida, M. *et al.* Genome sequencing and analysis of *Aspergillus oryzae*. *Nature* **438**, 1157–1161 (2005).
16. Pel, H.J. *et al.* Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. *Nat. Biotechnol.* **25**, 221–231 (2007).
17. Díez, B. *et al.* The cluster of penicillin biosynthetic genes. Identification and characterization of the *pcbAB* gene encoding the alpha-aminoadipyl-cysteinyl-valine synthetase and linkage to the *pcbC* and *penDE* genes. *J. Biol. Chem.* **265**, 16358–16365 (1990).
18. Merrien, M.A., Scott, P.M. & Polonsky, J. Roquefortine and isofumigaclavine A, alkaloids from *Penicillium roqueforti*. *Ann. Nutr. Aliment.* **31**, 963–968 (1977).
19. Kim, K.W. *et al.* Structure of malformin B, a phytotoxic metabolite produced by *Aspergillus niger*. *Biosci. Biotechnol. Biochem.* **57**, 787–791 (1993).
20. Gutiérrez, S., Fierro, F., Casqueiro, J. & Martín, J.F. Gene organization and plasticity of the beta-lactam genes in different filamentous fungi. *Antonie v. Leeuwenhoek* **75**, 81–94 (1999).
21. Fierro, F. *et al.* Transcriptional and bioinformatics analysis of the 58.6 kb DNA region amplified in tandem repeats containg the penicillinGene cluster in *Penicillium chrysogenum*. *Fungal Genet. Biol.* **43**, 618–629 (2006).
22. Van den Berg, M.A., Westerlaken, I., Leeflang, C., Kerkman, R. & Bovenberg, R.A.L. Functional characterization of the penicillin biosynthetic gene cluster of *Penicillium chrysogenum* Wisconsin54–1255. *Fungal Genet. Biol.* **44**, 830–844 (2007).
23. Lamas-Maceiras, M., Vaca, I., Rodríguez, E., Casqueiro, J. & Martín, J.F. Amplification and disruption of the phenylacetyl-CoA ligase gene of *Penicillium chrysogenum* encoding an aryl-capping enzyme that supplies phenylacetic acid to the isopenicillinN-acyltransferase. *Biochem. J.* **395**, 147–155 (2006).
24. Wang, F.Q. *et al.* Molecular cloning and functional identification of a novel phenylacetyl-CoA ligase gene from *Penicillium chrysogenum*. *Biochem. Biophys. Res. Commun.* **360**, 453–458 (2007).
25. Kovacevic, S., Tobin, M.B. & Miller, J.R. The beta-lactam biosynthesis genes for isopenicillinN epimerase and deacetoxycephalosporin C synthetase are expressed from a single transcript in *Streptomyces clavuligerus*. *J. Bacteriol.* **172**, 3952–3958 (1990).
26. Ullán, R.V., Casqueiro, J., Bañuelos, O., Fernández, F.J., Gutiérrez, S. & Martín, J.F. A novel epimerization system in fungal metabolism involved in the conversion of isopenicillinN into penicillin in *Acremonium chrysogenum*. *J. Biol. Chem.* **277**, 46216–46225 (2002).
27. Ullan, R.V., Campoy, S., Casqueiro, J., Fernandez, F.J. & Martin, J.F. Deacetylcephalosporin C production in *Penicillium chrysogenum* by expression of the isopenicillinN epimerization, ring expansion, and acetylation genes. *Chem. Biol.* **14**, 329–339 (2007).
28. Muller, W.H. *et al.* Localization of the pathway of the penicillin biosynthesis in *Penicillium chrysogenum*. *EMBO J.* **10**, 489–495 (1991).
29. Gidijala, L., van der Klei, I.J., Veenhuis, M. & Kiel, J.A.K.W. Reprogramming of *Hansenula polymorpha* for penicillin production: expression of the *Penicillium chrysogenum pcl* gene. *FEMS Yeast Res.* **7**, 1160–1167 (2007).
30. Kiel, J.A.K.W., van der Klei, I.J., van den Berg, M.A., Bovenberg, R.A.L. & Veenhuis, M. Overproduction of the peroxin Pex11p is associated with enhanced penicillin production *by Penicillium chrysogenum*. *Fungal Genet. Biol.* **42**, 154–164 (2005).
31. Baerends, R.J. *et al.* Genome2D: a visualization tool for the rapid analysis of bacterial transcriptome data. *Genome Biol.* **5**, R37 (2004).
32. Kunau, W.-H. (ed.). Peroxisomes: morphology, function, biogenesis and disorders. *Biochim. Biophys Acta* **1763** (2006).
33. Harris, D.M. *et al.* Enzymic analysis of NADPH metabolism in beta-lactam-producing *Penicillium chrysogenum*: presence of a mitochondrial NADPH dehydrogenase. *Metab. Eng.* **8**, 91–101 (2006).
34. Piper, M.D. *et al.* Reproducibility of oligonucleotide microarray transcriptome analyses. An interlaboratory comparison using chemostat cultures of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **277**, 37001–37008 (2002).
35. Gordon, M., Pan, S.C., Virgona, A. & Numerof, P. Biosynthesis of penicillin. I. Role of phenylacetic acid. *Science* **118**, 43 (1953).
36. MacQueen, J. Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. Math. Stat. Probability* **1**, 281–297, (Univ. California Press, Berkeley, 1967).
37. Rodríguez-Sáiz, M. *et al.* Reduced function of a phenylacetate-oxidizing cytochrome P450 caused strong genetic improvement in early phylogeny of penicillin-producing strains. *J. Bacteriol.* **183**, 5465–5471 (2001).
38. Martín, J.F. Molecular control of expression of penicillin biosynthesis genes in fungi: regulatory proteins interact with a bidirectional promoter region. *J. Bacteriol.* **182**, 2355–2362 (2000).
39. Kosalková, K. *et al.* A novel heptameric sequence is the binding site for a protein required for high level expression of *pcbAB*, the first gene of the penicillin biosynthesis in *Penicillium chrysogenum*. *J. Biol. Chem.* **275**, 2423–2430 (2000).
40. Bok, J.W. & Keller, N. LaeA, a regulator of secondary metabolism in *Aspergillus* spp. *Eukaryotic Cell* **3**, 527–535 (2004).
41. Van den Berg, M.A. *et al.* Method for enhancing secretion of beta-lactam transport. PCT patent WO 2001/32904 (2001).
42. Andrade, A.C., Van Nistelrooy, J.G., Peery, R.B., Skatrud, P.L. & De Waard, M.A. The role of ABC transporters from *Aspergillus nidulans* in protection against cytotoxic agents and in antibiotic production. *Mol. Gen. Genet.* **263**, 966–977 (2000).
43. Forster, J., Famili, I., Palsson, B.O. & Nielsen, J. Large-scale evaluation of *in silico* gene deletions in *Saccharomyces cerevisiae*. *OMICS* **7**, 193–202 (2003).
44. Bailey, J.E. *et al.* Inverse metabolic engineering: a strategy for directed genetic engineering of useful phenotypes. *Biotechnol. Bioeng.* **79**, 568–579 (2002).
45. Xu, Z. *et al.* Genome physical mapping from large-insert clones by fingerprint analysis with capillary electrophoresis: a robust physical map of *Penicillium chrysogenum*. *Nucleic Acids Res.* **33**, e50 (2005).
46. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
47. Salamov, A.A. & Solovyev, V.V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
48. Frishman, D. *et al.* Functional and structural genomics using PEDANT. *Bioinformatics* **17**, 44–57 (2001).
49. Cuomo, C.A. *et al.* The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science* **317**, 1400–1402 (2007).
50. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).